CS146 - Statistics

Final Project

Anna Pauxberger

21 December 2018

Motivation & Scenario

Carbon dioxide (CO_2) levels are rapidly increasing, contributing to climate change. While the source of climate change and its future implications are debated, the increase of CO_2 level increase is undoubted. How fast it is increasing, and at what time it will reach worrisome levels will be discussed in the following report. High levels of CO_2 impose health risks for humans, and depending on its urgency governments may want to proactively implement policies to limit its increase.

The following report uses pmm levels measured at Mauna Loa from 1958-2018 from the Scripps CO_2 program dataset. Using Bayesian inference techniques and Python's Stan package, data will be modeled until 2018, and predicted until 2058.

Assumptions

For the data, it is assumed that there was no change in technology or measurement that could have contributed to the change in levels. It assumes that the long term time trend is quadratic, and that seasonal variations can be modeled by a trigonometric function (specified below). It accounts for potentially missing data by modeling the days between observations, rather than just observations and assuming equal distance. The model itself assumes that variation is normally distributed.

The Model

Observed: CO_2 levels at time t

Unobserved: parameters for time trend (constant c_0, linear term c_1, quadratic term c_2), parameters for seasonal changes (amplitude c_3, c_4, phase c_5), noise (sigma)

$$p(x_t|\theta) = N(c_0 + c_1 t + c_2 t^2 + c_3 \frac{1}{c_4} \arctan(\frac{c_4 \sin(\frac{2\pi t}{365.25}) + c_5}{1 - c_4 \cos(\frac{2\pi t}{365.25}) + c_5}), \sigma)$$

Priors

Priors are defined according to what values they are most likely to take on, see meaning column in table below. However, since there is so much data available, most priors will not have a huge impact on the posteriors, since the likelihood of the data will get stronger as the amount of data points increases.

Variables for time trend

variable	limits	prior	meaning
$c0$	> 0	Normal(310,20)	The constant determines the intercept, and should be somewhere close to 300, which was approximately the level in 1958.
$c1$	> 0	Normal(1,2)	The linear term is difficult to estimate as it might be offset by the quadratic term, but it is positive and should not be larger than 4, as that would be too steep.
$c2$	> 0	Normal(1,2)	Similarly, the quadratic term is difficult to estimate and unknown to me. Therefore, the prior is large, centered on 1.

Variables for seasonal changes

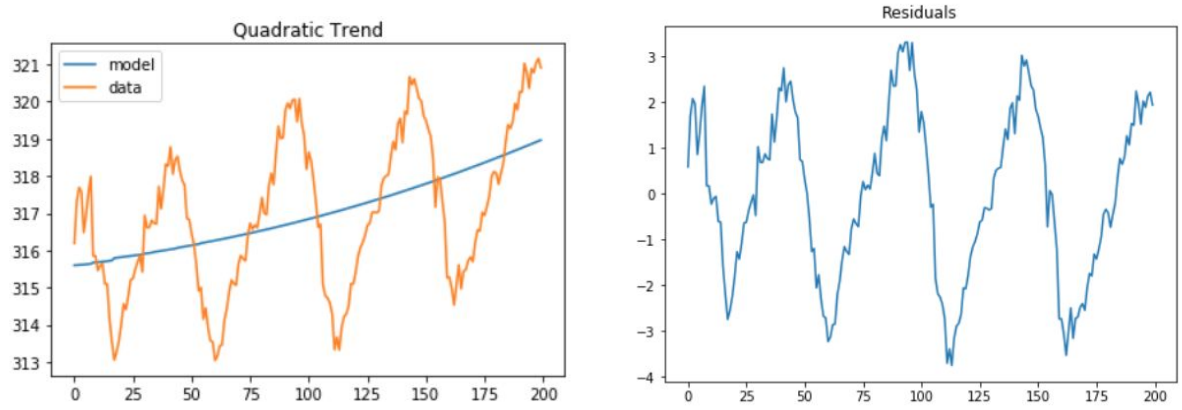
Experimented with on Desmos (https://www.desmos.com/calculator) using the equation

$$p(x_t|\theta) = N(c_0 + c_1 t + c_2 t^2 + c_3 \frac{1}{c_4} \arctan(\frac{c_4 \sin(\frac{2\pi t}{365.25}) + c_5}{1 - c_4 \cos(\frac{2\pi t}{365.25}) + c_5}), \sigma)$$

variable	limits	prior	meaning
$c3$	> 0	Normal(3,2)	The amplitude changes the y-range of the function. It should be positive since it otherwise flips, and below 5, as it otherwise would be too steep.
$c4$	$-1 < c4 > 0$	Cauchy(-0.5,1)	If this parameter is outside (-1,1) it is not continuous, and flips to be left tilted between 0 and 1. Since I don't know more about it, I set a broad prior centered on -0.5.
$c5$	<0	Normal(0,1)	Prior knowledge about the phase is difficult to incorporate, thus a rather broad prior is chosen for values around 1.
sigma $\sigma\sigma$	> 0	Normal(1.5,1)	The standard deviation models the noise we can have in our data. An estimate for a day could be for example 365.5 ppm, but probably can vary on average by 1-2 ppms for a given observation. I don't expect the variation to be too large as the year by year change itself is not too large.

How I arrived at the functional form

I first experimented with the first 200 data points (~4 years) to test different time trends (linear, quadratic, exponential) as well as seasonal variations (sin, cos, variations thereof). When I found a version that looked good, I extended it to a larger time span. This sometimes lead to complications, as certain variations might not have been accounted for by the previous model. To arrive at the correct functional form of the model, I modeled the time trend only first, and then plotted the residuals to see a pattern. In that pattern I could identify that the function was decreasing more steeply then increasing, looking like a tilted sharp sine function. A helpful source online as well as a desmos calculator helped me to arrive at the above specified seasonal variation.



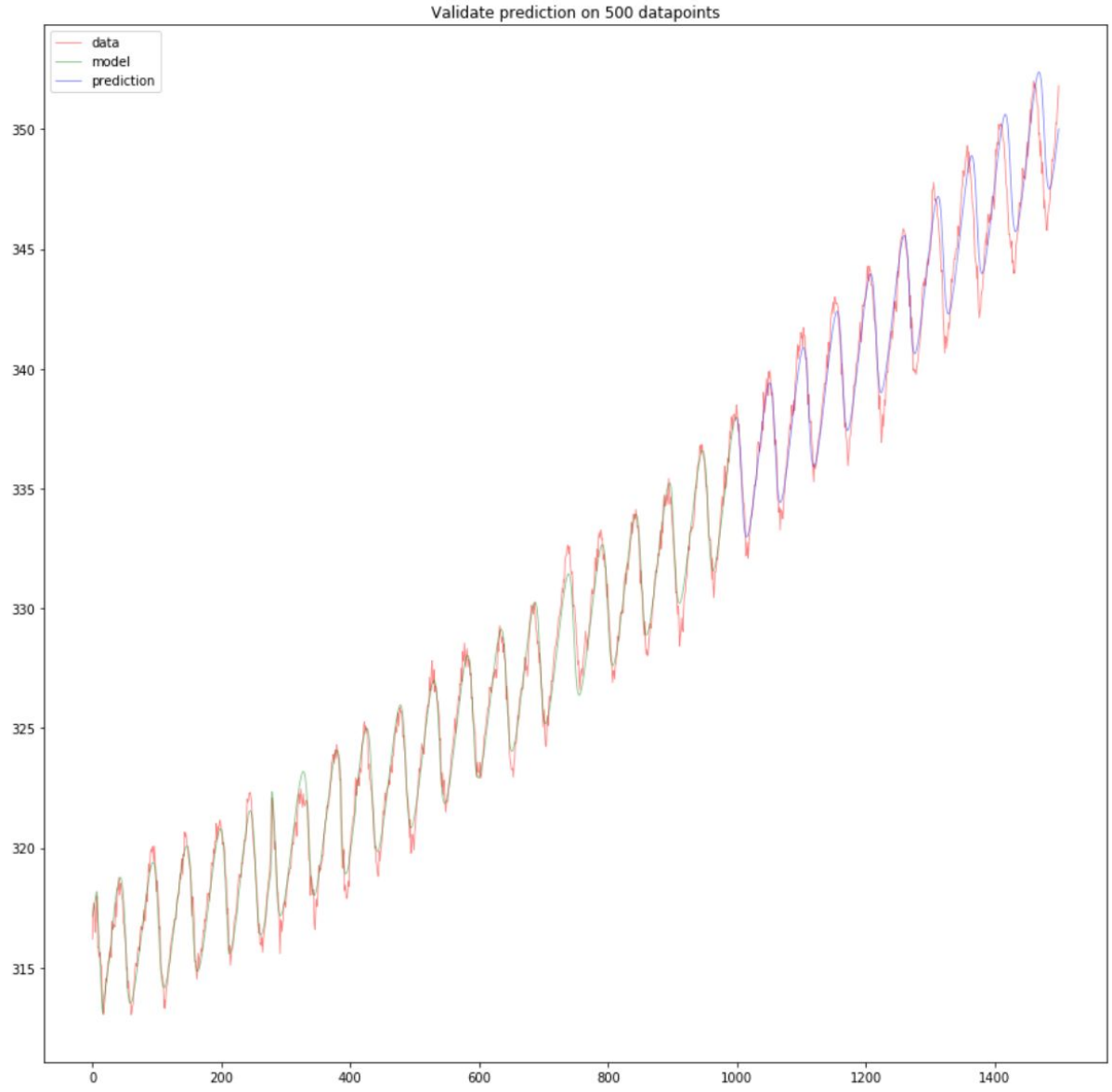
Stan Model Performance

The model performed with Rhats values of 1 for each variable, suggesting that the sampled chains mixed well and the variance between chains equals the variance within chains. The effective sample size (n_eff) is above 1000, out of 4000 samples), for each variable, which is plenty to have believable results. That still means that around 75% of samples were correlated, but as long as we have enough independent samples themselves it is a matter of computational efficiency and not results.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
c_0	314.5	7.7e-4	0.05	314.41	314.47	314.5	314.53	314.58	3575	1.0
c_1	2.1e-3	1.8e-7	9.6e-6	2.1e-3	2.1e-3	2.1e-3	2.1e-3	2.1e-3	2890	1.0
c_2	9.6e-8	7.7e-12	4.2e-10	9.5e-8	9.5e-8	9.6e-8	9.6e-8	9.7e-8	2937	1.0
c_3	2.84	5.8e-4	0.02	2.8	2.83	2.84	2.85	2.88	1279	1.0
c_4	-0.5	3.2e-5	1.5e-3	-0.51	-0.5	-0.5	-0.5	-0.5	2169	1.0
c_5	1.12	1.6e-4	6.6e-3	1.11	1.12	1.12	1.12	1.13	1681	1.0
sigma	0.84	1.8e-4	0.01	0.82	0.83	0.84	0.84	0.86	3660	1.0

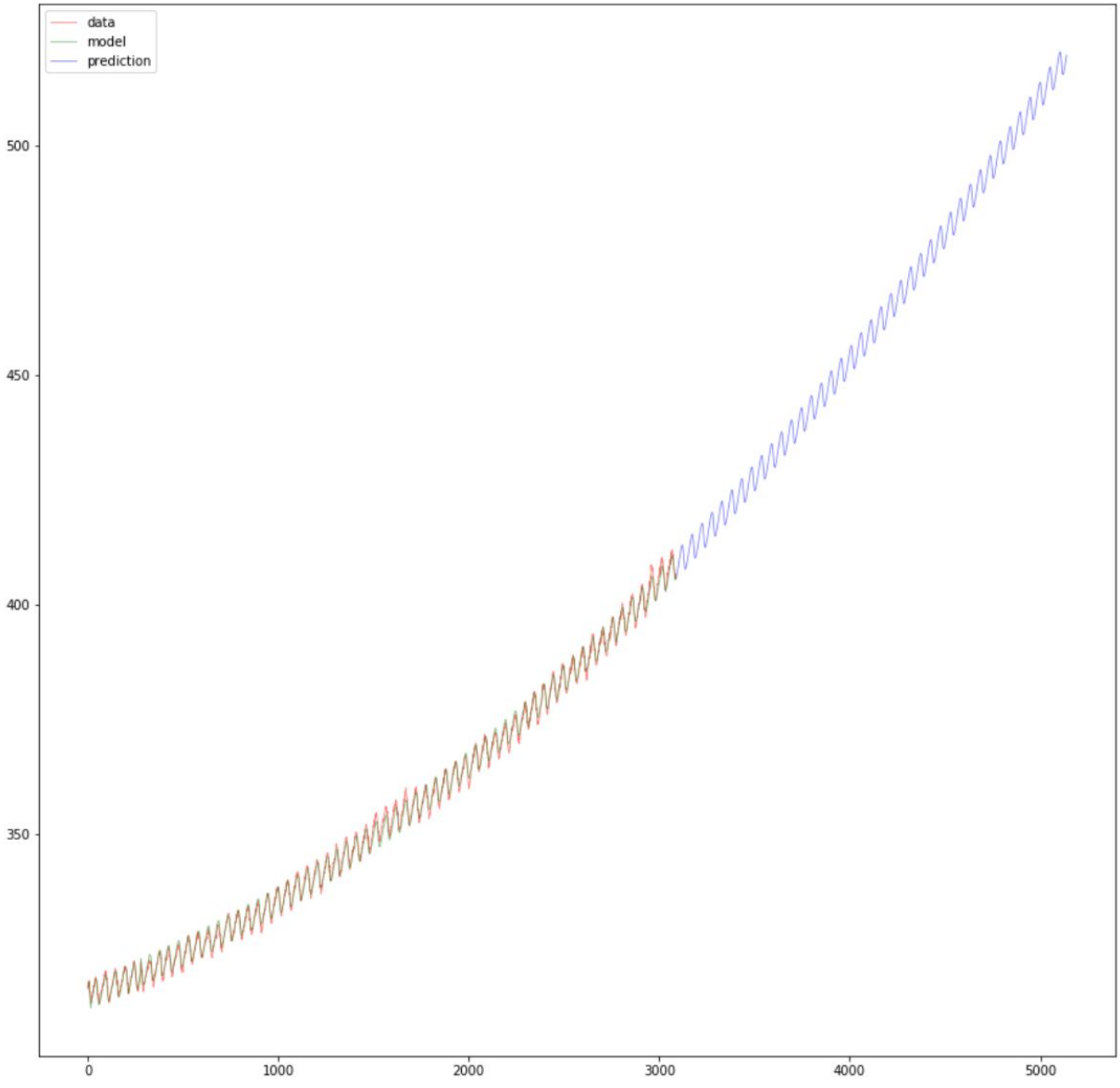
Validation Set

Before training the model on the entire dataset, I provide 1000 data points, and predict 500. The figure below shows that the predictions lie almost perfectly on the data, indicating a well fit model. Narrow troughs/ peaks are not well captured by the model however, meaning it over,-under-estimates CO2 levels during winter and summer times - something the model can be improved upon.



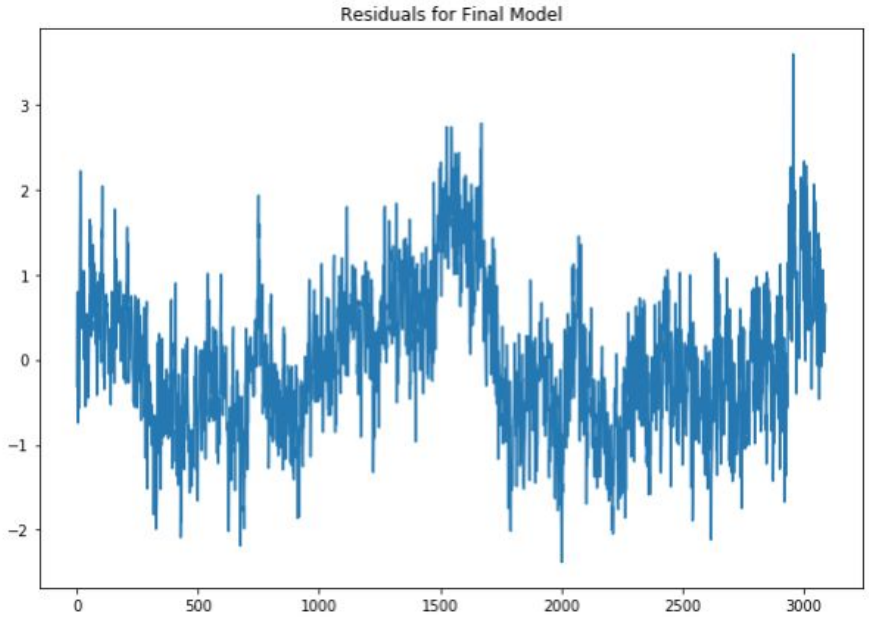
Model Performance on Entire Dataset

Even when trained on the entire dataset, with a range of 1958-2018, the model performs reasonably well.



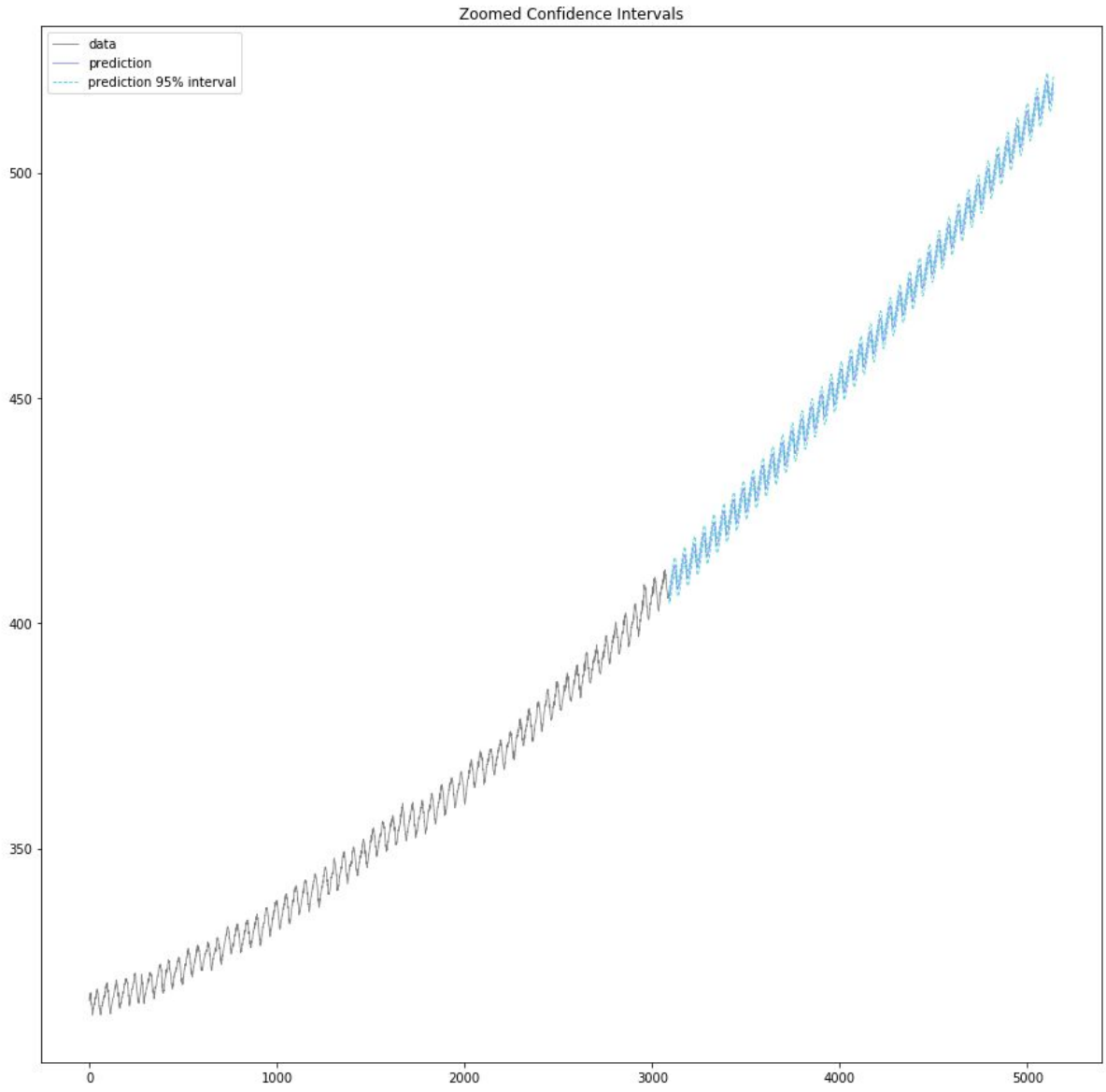
Errors

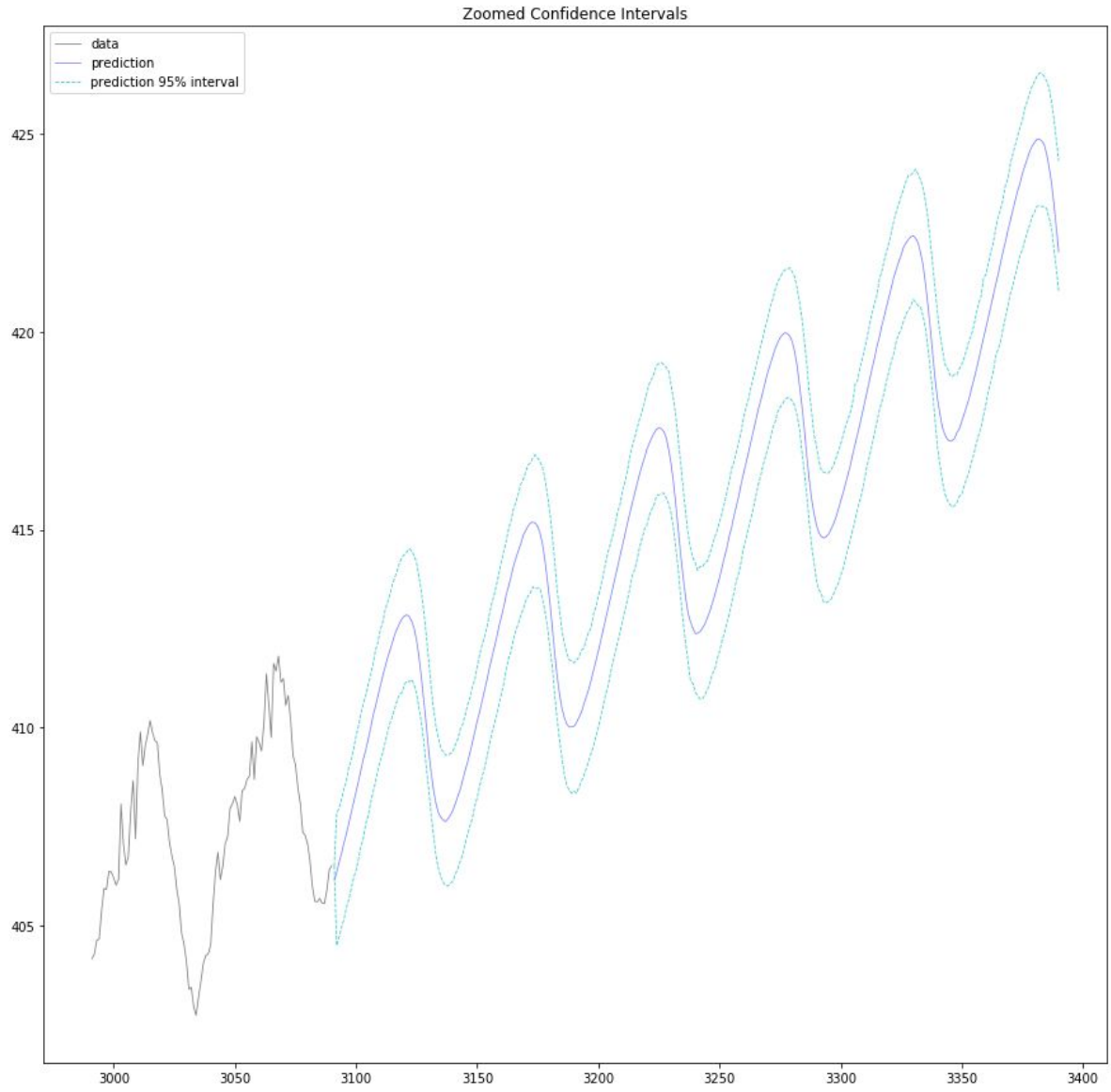
The residuals between the prediction and the error seem to have a wave shape to it, which could be included in the model to improve prediction accuracy. There might be a CO2 behavior that climate experts know of who could help figure out the behavior of the residuals.



Confidence Intervals

The narrow confidence intervals for predictions up to 40 years into the future first seem unreasonable. However, when looking at the validation set, the predictions from the model proved themselves to be very accurate, thus it can allow a narrow confidence interval. If the functional form, however, is misspecified for the data we have available and is in fact steeper or flatter, thos predictions can easily be off-set.

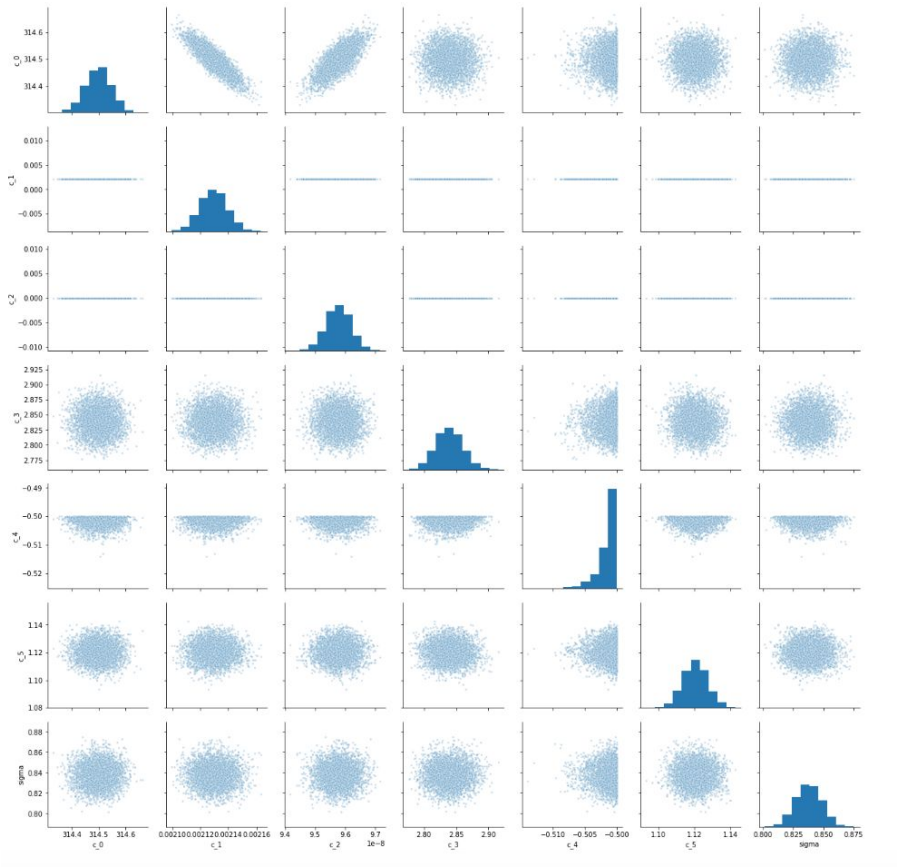

Zoomed Confidence Intervals
data
prediction
prediction 95% interval
500
450
400
350
0
1000
2000
3000
4000
5000


Zoomed Confidence Intervals
data
prediction
prediction 95% interval
425
420
415
410
405
3000
3050
3100
3150
3200
3250
3300
3350
3400

Model Assessment

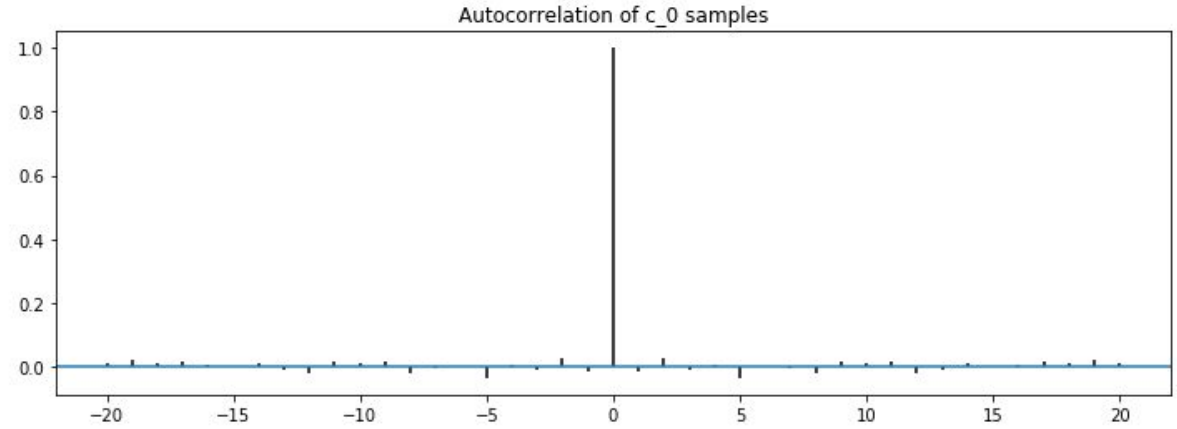
Pairplots

In the pairplot graph below we can see that all parameters are normally distributed, with the exception of c_4, which was restricted to be between -0.5 and -1 and thus portrays a truncated distribution. Furthermore, the constant c_0 is negatively correlated with the linear term c_1 and positively correlated with the quadratic term c_2. The negative relationship between c_0 and c_1 is reasonable, as a higher intercept models the data better with a lower linear incline. I would not have expected the constant to be positively correlated with the quadratic term, but apparently, a higher intercept models the data better with a higher quadratic term.



Autocorrelation Graphs

The autocorrelation graph shows that there is almost no autocorrelation among the samples, indicating that the chains mixed well and we have a rather high level of independent samples. This is also visible in the output above, where the effective sample size is above 1000 in the worst case (c_4), and well above 2000 in most cases. (The entire output for all variables can be found in the Notebook. They are omitted here as they look very similar, and sho no sign of autocorrelation.)



Prediction for January 2058: 519ppm CO2

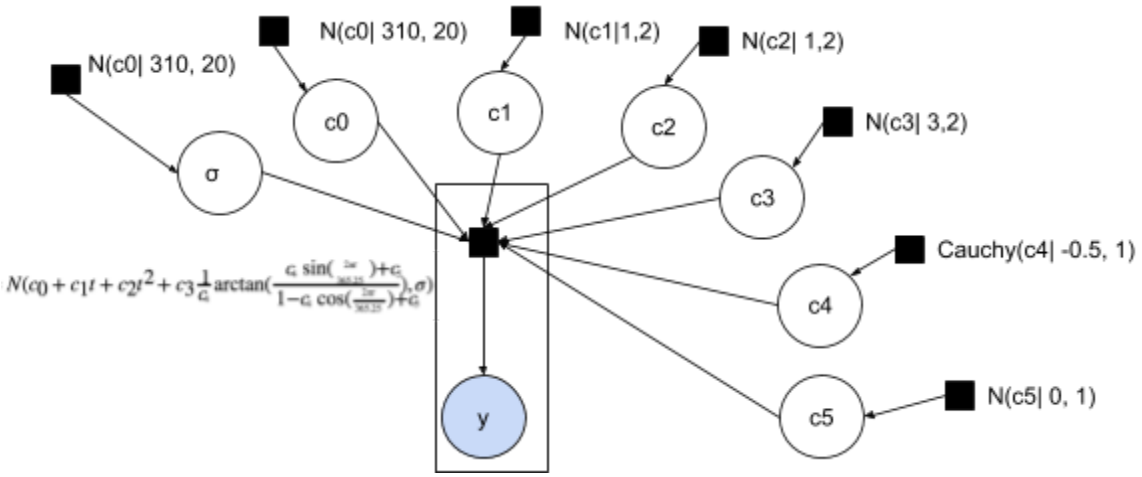
This model predicts CO2 (ppm) levels for January 2058 to be at 519.59 with a 95% confidence interval of 517.85 and 521.31.

High Risk starting Dec 2032: 450ppm

Continuing with the same rate CO2 levels are currently rising, we can be 95% confident that will observe a CO2 level between 448 and 451 CO2 ppm on 11 December 2032, which is in 14 years. Health risks that high can cause tremendous problems for the health of the populations, the environment and governments' policies.

Directed Graph

The directed graph below shows how the observation y (CO2 ppm) is generated from its priors and the associated factors.



Complications

Normalizing the CO_2 level data proved not to enhance the model, but complicated the denormalization for the prediction. Therefore, the CO_2 level data remained un-normalized.

Modeling the seasonal variation was a challenge, as its behavior is not easily visible, and difficult to model. In the end, it helped to model just the quadratic time trend alone, take a look at the behavior of the residuals and then identify a tilted function, which I found a solution to via online search (trigonometric function). Choosing the correct **priors**, especially for the seasonal variation, was difficult, as the parameters for the trigonometric function are not easily intuitively understandable and my prior knowledge on climate data as well as the implementation of the parameters is limited. For example, I thought that I could use a prior of a uniform distribution to model the phase, which turned out worse than a normal centered on 1.

Shortcomings

There is a pattern of the model under and over predicting in a wave form. It would be interesting including knowledge about century variations in the model as well. The current data seems to fit a quadratic function well, however, given the data is only provided for the past 60 years, another time trend function might fit the long term trend better and make more accurate predictions.

Conclusion

The predictions from the specified model have been proven robust and accurate when compared to a validation set and when tested on Stan's performance (Rhat, n_eff, residuals). With narrow 95% confidence intervals, the predictions are fairly accurate, unless an unexpected change will cause the quadratic time trend to shift.